NIPES

Journal of Science and Technology Research

Journal homepage: www.nipesjournals.org.ng

A Novel Model Developed for Forecasting Oilfield Production Using Multivariate Linear Regression Method

Chukwu Emek,[a*] and Kelani Bello[b]

[a,b] Department of Petroleum Engineering, University of Benin, Benin City, Edo State Nigeria

* Email: pastorechukwu@gmail.com

ARTICLE INFORMATION

Article history:
Received 28 April 2019
Revised 10 May 2019
Accepted 16 May 2019
Available online 06 June 2019

Keywords:
Forecasting oilfield production, model, Group Gatheing Facility, Multivariate Linear Regression

ABSTRACT

In this paper, a multivariate linear regression model was developed for predicting crude oil production volume in a group gathering facility within the Niger delta area of Nigeria.
The dataset used was split randomly into two parts namely the training and testing data set. This was done to ensure the model was not over fitted. The model depends on four (4) independent variables: volume correction factor, metered volume, metered factor and gross standard volume for accurate predictions of net oil volume (dependent variable). The model was compared with other existing models and was found to be more accurate and has better performance in terms of root mean square error and residuals. This novel model is suggested for use by the oilfield managers to assist in decision making.

1. Introduction

Crude oil exploitation has become the mainstay of many economies throughout the world. Nigeria, for example, has its economy greatly dependent on revenues from oil and gas business. The World Bank reported that in Nigeria, over 95% of export income and 85% of government proceeds come from oil [1]. Accurate prediction of oilfield output is important for planning especially in countries whose economies are highly dependent on oil [2]. The world's energy needs are largely catered for by crude oil production, accurate forecast is therefore desirable. The field managers saddled with the responsibility of planning and decision making, relies heavily on oil production volume [2]. The prediction of the output of oilfield can be done by artificial neural networks, grey prediction method, logistic curve method, fuzzy logic, regression, curve fitting, Weng cycle model, etc. with each of these having their own applicability and limits.

Researchers have conducted an investigation into forecasting of oil reserves and production in oil fields. Wu Xin-gen [3] applied the artificial neural network in predicting the output of oil fields and found that the artificial neural network (ANN) was a feasible predicting method after he compared the results from ANN to Weng cycle model. Huang et al developed a new model for oilfield performance. Their model was based on the one-way principle. The emulation calculation indicated that the prediction accuracy of the model was satisfactory.

Wang et al [4] stated that oilfield exploitation is a complicated multivariate non-linear dynamic system. In their paper, they optimized a multivariate oilfield output prediction model by using

multivariate linear regression and ANN. They found from their study that the optimized model is simpler and more useful and it can make the prediction precise with fewer sample data.

Senan et al [5] designed the fuzzy petroleum prediction as an expert system and used five petroleum production factors which were temperature, pressure, crude oil density, gravity, and gas density. They reported results close to empirical values. Chen et al [6] published a new efficient model based on the lognormal distribution in probability statistics.

The multilinear regression has been used for oilfield output prediction in some oilfields outside Nigeria. Izni and Radzuan [7] used the multilinear regression while Na et al [8] used the least square fitting method.

Furthermore, the use of support vector regression (SVR) and ANN to predict oilfield production complies with the actual oilfield production dynamics and the prediction error of them are less than 10 %. However, they being learning models imply the need for a large dataset for training and prediction, thus making them suitable for the short-term prediction. Other methods can be used to see the trend in oilfield production although such models may not be very accurate. An example of such models is the GM (1,1) predicting model.

The multivariate normal regression model has three main advantages that make it suitable for use in this study. The first is the ability to determine the relative influence of one or more predictor variables to the criterion value. The second advantage is the ability to identify outliers, or anomalies, and the third is: it's generally high accuracy.

In this study, a highly accurate predictive model has been developed for forecasting oilfield output using multivariate normal regression.

2.1. Methodology

The research was carried out by using MATLAB to analyze the true data set acquired from the oilfield (see Table 1). A new predictive model was then formulated from the multivariate analysis of the dataset.

The multivariate normal regression analysis method was used as the learning algorithm to which the datasets were fed to generate the models. The multivariate normal regression algorithm has the capacity to house several variables with potential high level of accuracy, which, unlike other modeling tools that can only analyze limited numbers of dimensions. This makes it a good fit for the prediction problem.

2.2. Multivariate models

Multivariate normal regression is the regression of a d – dimensional response on a design matrix of predictors with normally distributed errors. The errors can be differently dispersed and correlated.

The model is given by Equation (1) [9]:

$$y_i = X_i\beta + e_i\,, \qquad i = 1, \dots, n, \tag{1}$$

Furthermore, the expectation/conditional maximization and covariance-weighted least squares estimation algorithms include imputation of the missing response values.

Let $\tilde{y}$ represent the missing observations. Then conditionally imputed values are the expected value of the missing observation given the observed data, $E(\tilde{y}|y)$.

The joint distribution of the missing and observed responses is a multivariate normal distribution given in Equation (2),

$$\begin{pmatrix}\tilde{y}\\ y\end{pmatrix} \backsim MVN\left\{\begin{pmatrix}\widetilde{x_\beta}\\ x_\beta\end{pmatrix}, \begin{pmatrix}\Sigma_{\tilde{y}} & \Sigma_{\tilde{y}y}\\ \Sigma_{y\tilde{y}} & \Sigma_y\end{pmatrix}\right\} \quad (2)$$

By applying the properties of the multivariate normal distribution, the conditional expectation is as presented in Equation (3)

$$E(\tilde{y}|y) = \tilde{X}\beta + \Sigma_{\tilde{y}y}{\Sigma_y}^{-1}(y - X\beta) \quad (3)$$

It should be noted that the function only imputes missing response values. Observations with missing values in the design matrix are however removed.

2.3 Model Performance Parameters

The model performance parameters are calculated from Equations (4-7)

The root mean square error (RMSE) score can be calculated from Equation (4):

$$RMSE = \sqrt{\frac{\Sigma(Predicted_i - Observed_i)^2}{Number\ of\ Examples}} \quad (4)$$

$$Residual = Observed_i - Predicted_i \quad (5)$$

Proportional reduction of error (PRE)

$$PRE = \frac{Residual\ Sum\ of\ Squares\ of\ Model\ 2 - Residual\ Sum\ of\ Squares\ of\ Model\ 1}{Residual\ Sum\ of\ Squares\ of\ Model\ 2} \quad (6)$$

$$Residual\ Sum\ of\ Squares\ of\ Model = \sum_{i=1}^{n}(y_i - f(x_i))^2 \quad (7)$$

$$R^2 = 1 - \frac{SSE}{SST} \quad (8)$$

2.4 Existing Model

$$Qn = \mathrm{Q\,i} \times \mathrm{M\,F} \times \mathrm{VCF} \times \mathrm{BS\&}W \quad (9)$$

Equation 8 is the model currently in use at the Umusadege GGF

$$BSW = 1 - \left(\frac{\mathrm{BS\&W\ \%}}{100}\right) \quad (10)$$

2.5 Data Used

The Data was acquired from an oilfield gathering facility in Kwale, Niger Delta, Nigeria. It comprised of: net oil volume (the target variable), while the predictors included: metered factor, volume correction factor, the basic sediment and water, metered volume, the API @60, gross standard volume, temperature etc.

Table 1: Data from oilfield gathering facility in Kwale, Niger Delta, Nigeria

Months	Gross Vol	API	Temp	Metered Vol	MF	VCF	BS&W	Net
Feb-09	2087	42.1	116.3	7110	1.0003	0.9707	0.3	6883.036
Mar-09	2966	42.1	116.1	6615	1.0003	0.971	0.3	6405.817

Apr-09	2864	41.9	115.4	6190	1.0003	0.9713	0.3	5996.108
May-09	1989	42.3	116.1	7273	1.0003	0.9708	0.35	7038.027
Aug-09	2747	41.9	116.9	6678	1.0003	0.9705	0.3	6463.494
Jul-10	2445	42	117.5	6143	1.0003	0.9702	0.25	5946.822
Aug-10	2689	42	117.8	6027	1.0003	0.9699	0.25	5832.723
Sep-10	2163	42	118.1	6386	1.0003	0.9699	0.2	6183.248
Oct-10	3127	42.2	114.7	6622	1.0003	0.9715	0.25	6419.115
Nov-10	2811	42.4	114.6	7117	1.0003	0.9716	0.2	6903.118
Dec-10	2366	42	114.6	7205	1.0003	0.9718	0.2	6989.912
Jan-11	2037	43.4	115.2	7003	1.0003	0.971	0.2	6788.349
Feb-11	2813	42	118.1	6250	1.0003	0.9699	0.35	6042.471
Mar-11	2778	42.1	119.3	6237	1.0003	0.9692	0.25	6031.597
Apr-11	2460	41.8	119.4	6411	1.0003	0.9692	0.2	6202.974
May-11	2830	41.7	118.8	6579	1.0003	0.9696	0.2	6368.15
Jun-11	2313	41.5	118.1	7642	1.0003	0.9701	0.5	7378.65
Jul-11	2158	41.6	117.9	6955	1.0003	0.9701	0.37	6724.098
Aug-11	1186	41.6	116.8	7531	1.0003	0.9706	0.2	7297.158
Sep-11	2447	41.7	119.1	6064	1.0003	0.9696	0.2	5869.655
Oct-11	2181	42.1	118.4	7183	1.0003	0.9697	0.2	6953.51
Nov-11	2082	41.9	118.3	7097	1.0003	0.9697	0.2	6870.257
Dec-11	1815	41.8	120.3	6391	1.0003	0.9686	0.25	6176.699
Jan-12	2444	42	118.8	6440	1.0003	0.9694	0.25	6229.197
Feb-12	2682	41.5	116.9	6446	1.0003	0.9706	0.3	6239.589
Mar-12	1268	41.3	117.3	7024	1.0003	0.9704	0.3	6797.68
Apr-12	2680	41.2	118.5	7501	1.0003	0.97	0.3	7256.318
May-12	968	41.2	116.9	6410	1.0003	0.9706	0.25	6207.854
Jun-12	2642	41.4	117.2	6222	1.0003	0.9706	0.3	6022.762
Jul-12	2970	41.8	117.4	5944	1.0003	0.9702	0.3	5751.293
Aug-12	2509	42.2	117.8	6329	1.0003	0.9699	0.3	6121.918

2.5.1 Data Description

The multivariate normal regression algorithm functions by learning the relationship between the dimensions of the design matrix and connects this to the predicted variable. First, the data set is split into two parts of 70:30%. The learning is carried out on the first set of data (70%/training data) and the rest (30%/test data) is used to test the accuracy of the model. It should be noted here that the results of any predictive model should be tested with a portion of the data it did not see in the training phase as not doing this means that the model may not generalize in the real-world scenario. Furthermore, the issues of overfitting might arise if random data splitting was not carried out. Also, it is necessary to state here, that we used the root mean squared error to calculate the accuracy of the resultant model.

Table 2: Models Generated

	MODEL (X)	β
1	MV	0.98426
	MF	-2643.4

	VCF Temp BS&W API @ 60	2669.1 0.21421 -28.519 0.053712
2	MV MF VCF BS&W	0.98417 -2417.8 2461.9 -29.436
3	MV MF VCF Temp BS&W	0.98426 -2632.9 2662.3 0.20458 -28.591

2.6. A New Generalized Model

A new generalized model was formulated from the regression analysis carried out in this study, it was obtained by coefficient comparison and inspection of the model with the lowest root mean square error value (regression model 2) to arrive at a nicely looking model that involved the variables rather than have cumbersome numbers as coefficients: by inspection we can see that if we multiply the VCF and MV, we get the first term in the model, then if we multiply, GV and MF we get the second term, if we again multiply the VCF and MV, we get the third term in the model which can then be simplified as given by Equation 10. It is to be noted that the fourth term having BS&W because of it lower PRE (Figure1) when compared the remaining three independent variable.(Mv=1, MF=0.720,VCF=0.755).

$$Q_n = 2 \times VCF \times MV - MF \times GV \quad (10)$$

The existing model was dependent on the metered volume, metered factor, volume correction factor, and basic sediment and water percentage, while the new generalized model took the metered volume into consideration but in contrast to the existing model it neglected the basic sediment and water percentage. This will help the production manager make a quick projection while waiting for the analysis for basic sediment and water.

3. Results and Discussion

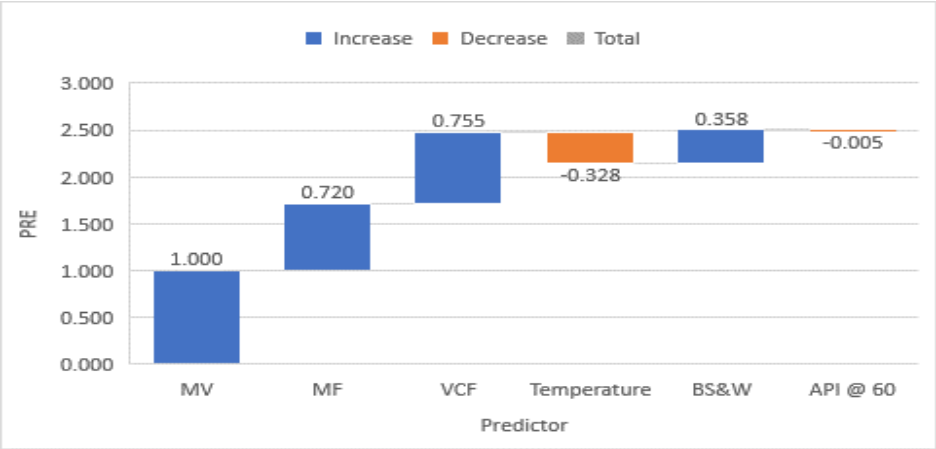

Figure 1: A Graph of Proportional Reduction of Error for the Predictors

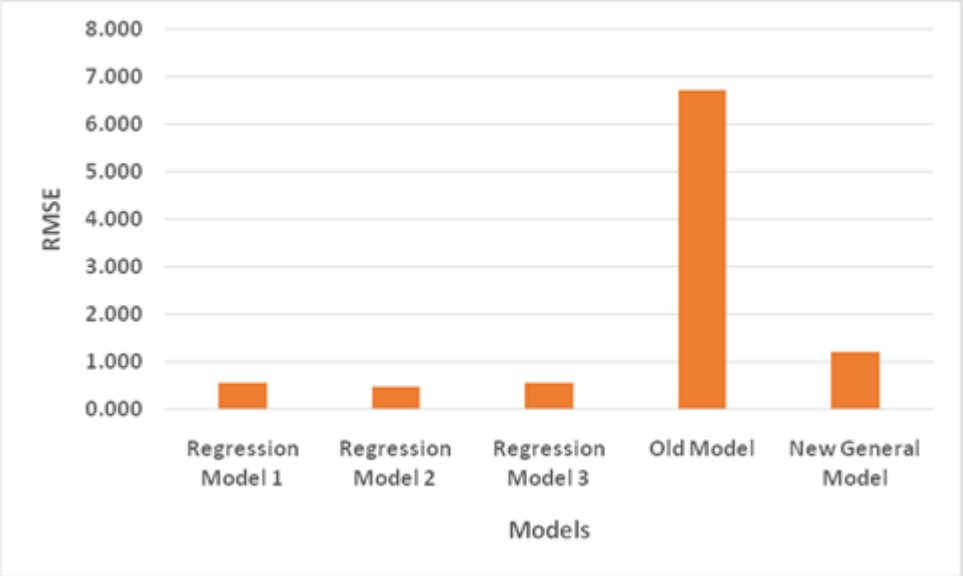

Figure 2: A Graph of the Root Mean Square Error of the Regression Models 1-3, Old Model and the New General Model

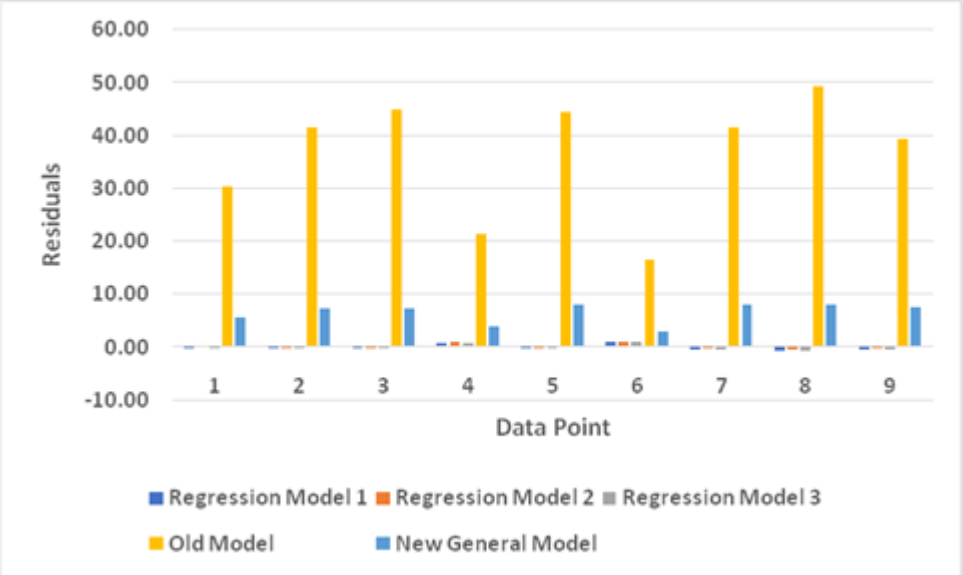

Figure 3: A Graph of the Residuals of the Regression Models 1-3, Old Model and the New General Model

As can be observed from Figure 1 and Figure 2 the proportional reduction of error, showed that the important predictors were the MV, MF, VCF, BS&W each having a proportional reduction of error value of 1, 0.72046, 0.75538, 0.35782. It is worthy of note that the proportional reduction of error represents the gain in precision of predicting a dependent variable from knowing the independent variable. In this case the gain in precision of using MV, MF, VCF, BS&W was significant while the other variables were not. The root mean square error is a commonly used measure of the differences between values predicted by a model and the values observed. The root mean square error serves to sum the magnitudes of the errors in predictions for various times into a single measure of predictive power. Root mean square error is a measure of accuracy, to compare forecasting errors of different models for a particular dataset and not between datasets.

This is scale-dependent. Root mean square error is always non-negative, and a value of 0 (almost never achieved in practice) would indicate a perfect fit to the data. In general, a lower root mean square error is better than a higher one. However, comparisons across different types of data would be invalid because the measure is dependent on the scale of the numbers used. In this study as shown in Figure 2, the root mean square error value was 0.49318 when model 2 (MV, MF, VCF, BS&W) was used while the root mean square error value of model 1 using all the six predictors was 0.56572. Furthermore, when the MV, MF, VCF, Temperature, and BS&W, were used as predictors the root mean square error value was 0.56439. it is worthy of note that these three configurations gave the lowest root mean square error values while the other configurations tested were considerably higher. Next the residuals were studied. A residual shows the difference between the predicted value and observed value. The lower it is for out-of-sample analysis the better the model. As can be observed from Figure 3, the residuals very high for the Model in equation 8 (the Old Model) and was satisfactorily low for the other models.
The generalized model was found to have a root mean square error value of only about 2.5 times higher than regression model 2 in this study whose root mean square error value was 0.49 and 6 times better than the root mean square error value of the model in Equation (9). The R-square for the new model looks good as it above 75% (Table 3).

Table 3: R^2 value of models

Model	R^2
Regression Model 1	0.99174
Regression Model 2	0.992
Regression Model 3	0.99177
Old Model	0.992
New General Model	0.98893

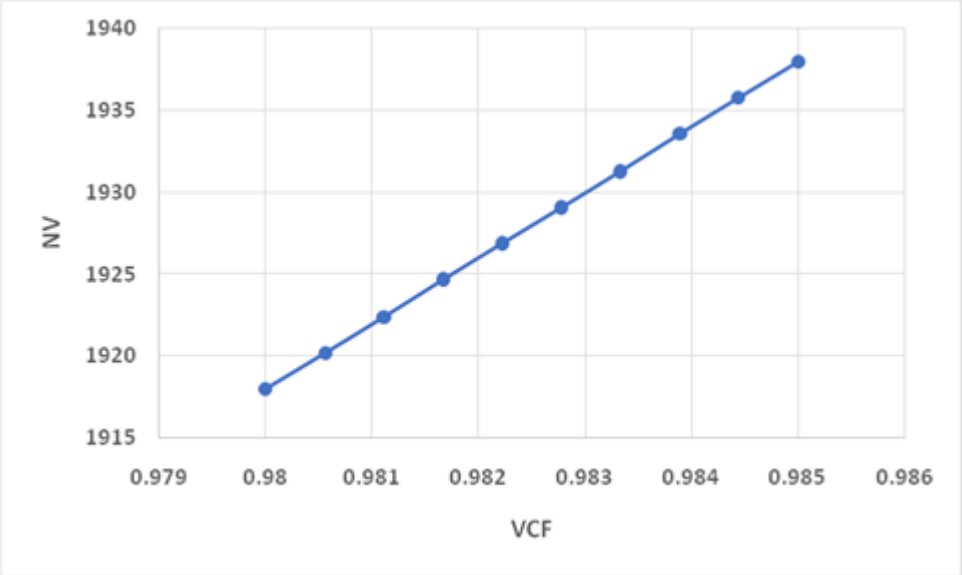

Figure 4: A Plot of the Net Standard Volume with the Volume Correction Factor

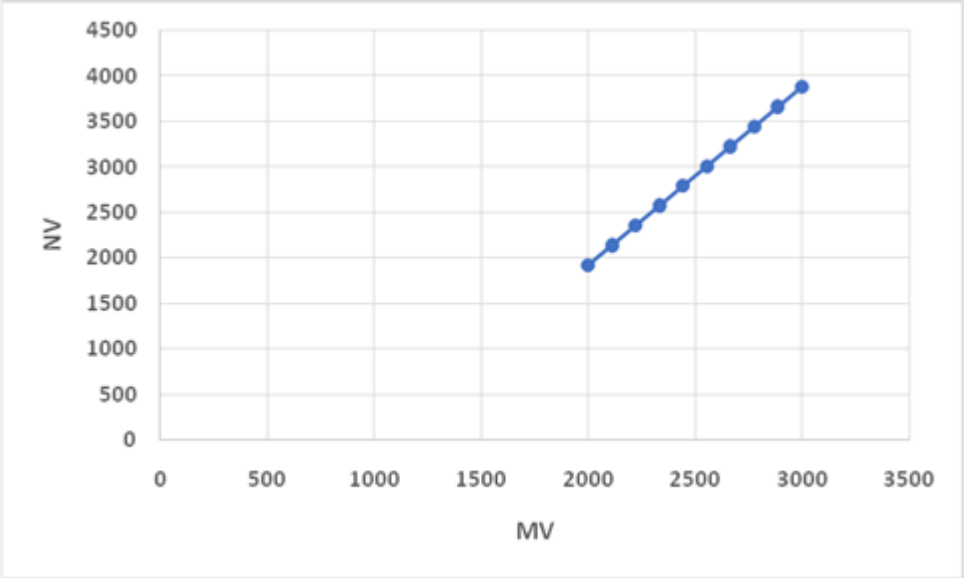

Figure 5: A Plot of the Net Standard Volume with the Metered Volume

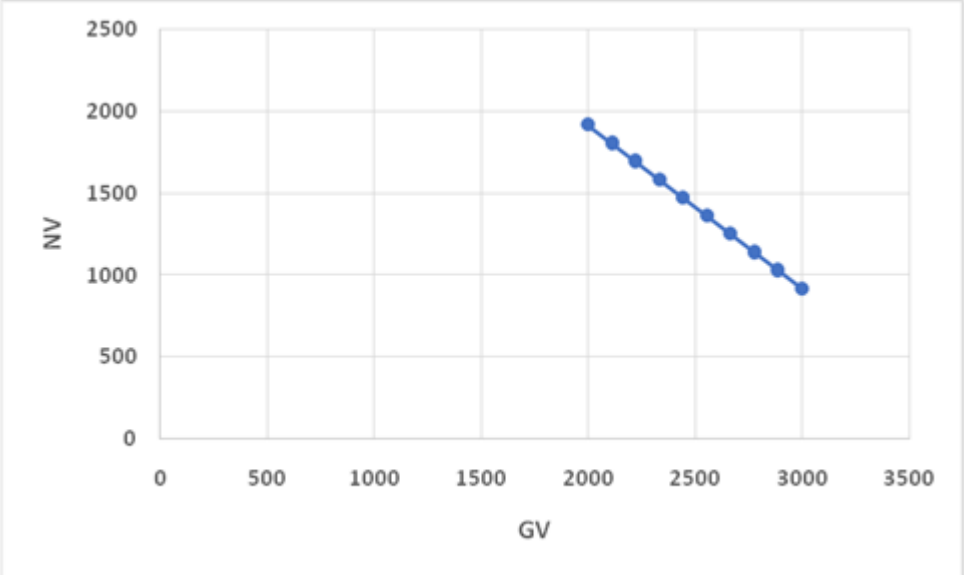

Figure 6: A Plot of the Net Standard Volume with the Gross Standard Volume

The predictor variables of the new general model vary linearly with the target variable. From Figure 4 we can observe a linear relationship of the volume correction factor (VCF) with the net standard volume (NV). The NV increases as VCF increases. Similarly, the net standard volume (NV) increases linearly with the metered volume (MV) as can be observed in Figure 5 However, the net standard volume (NV) decreases linearly with increase in gross standard volume (GV), this trend can be observed in Figure 6. Figures 4 – 6 agree with the assumption that there is a linear relationship between the dependent variable and the independent variable.

4. Conclusion

In this paper, a novel multiple linear regression model for predicting oilfield output has been developed. The data obtained was cleaned and fed into the MATLAB's regress routine, the routine learned the data and created a model from the data fed to it with varying number of predictors. Then the model was tested with the test data (30% of the whole dataset) to evaluate its accuracy.

The suggested model takes four (4) predictors to make its prediction. The model has been subjected to statistical analysis and has been found to be useful in oilfield output prediction. This model is hoped to be a very valuable tool in the hands of the oilfield manager for decision making in the oilfield.

Nomenclature

BS&W = basic Sediment and Water accounting for the fraction of water and contaminants, determined by sample analysis.

BS&W % = percentage of sediment and water measured

e_i = d – dimensional vector of error terms, with multivariate normal distribution

$GV = \text{Gross Standard Volume}$
MF = Meter Factor, adjust to actual volume, this factor is determined by proving the meter

$MV = \text{Metered Volume}$

VF = Volume Factor

Q_n = net oil volume

Q_i = gross measured oil volume (measured volume)

VCF = volume correction factor for the effects of Temperature and Pressure.

X_i = a design matrix of predictors

y_i = a d – dimensional vector of responses

Greek letters
β = vector or matrix of regression coefficients

References

[1] U. B. Akuru and O. I. Okoro (2011), "A prediction on Nigeria's oil depletion based on Hubbert's model and the need for renewable energy," *ISRN Renewable Energy*, vol. 3 pp22-27.
[2] M. Oladeinde, A. Ohwo, and C. Oladeinde(2015), "A Mathematical Model for Predicting Output in an Oilfield in the Niger Delta Area of Nigeria," *Nigerian Journal of Technology*, vol. 34, pp. 768-772.
[3] W. Xin-gen (1994), "The Application Of Artificial Neural Network In Predicting Output Of Oil Fields [J]," *Petroleum Exploration And Development*, vol. 3, pp34-37.
[4] T. Wang, x.-g. Chen, and Y.-f. li (2006), "Optimization of Multivariate Model in Oilfield Output Prediction [J]," *Computer Simulation*, vol. 2, p. 015.
[5] S. A. Ghallab, N. Badr, A. B. Salem, and M. Tolba (2013), "A Fuzzy expert system for petroleum prediction," *WSEAS, Croatia*, vol. 2, pp. 77-82.
[6] Y. Chen (1996), "Derivation and application of Weng's predication model," *Natural Gas Industry*, vol. 16, pp. 22-26.
[7] I. Izni binti Mustafar and R. Radzuan Razali (2011), "A study on prediction of output in oilfield using multiple linear regression," *International Journal of Applied Science and Technology*, vol. 1, pp. 107-113.
[8] W. B. Na, Z. W. Su, and P. Zhang (2013), "Research of oilfield production forecast based on least squares fitting and improved BP neural network," in *Applied Mechanics and Materials*, pp. 1456-1460.
[9] MathWorks. (R2006b, December 22). *mvregress*. Available: https://www.mathworks.com/help/stats/mvregress.html